

International Workshop on Web Search and Data Mining (WSDM) April 29 - May 2, 2019,
Leuven, Belgium

Integration of Data Mining Classification Techniques and Ensemble Learning for Predicting the Export Potential of a Company

Jesus Silva^{a*}, Jenny Romero Borré^b, Aurora Patricia Piñeres Castillo^c, Ligia Castro^d,
Noel Varela^e

^a Universidad Peruana de Ciencias Aplicadas, Lima 07001, Peru

^{b,c,d,e} Universidad de la Costa (CUC), Barranquilla 080003, Colombia

Abstract

In this research, data mining techniques are integrated with Ensemble Learning for predicting the export potential of a company. The analysis covers the stages of measurement, evaluation and classification of companies, based on a proposal of 16 key factors of the export potential. The techniques standing out are: Synthetic Minority Oversampling Technique (Smote), K-Means Clustering, Generalized Regression Neural Network (GRNN), Feed Forward Back Propagation Neural Network (FFBPN), Support Vector Machine (SVM), Decision Tree (DT) and Naive Bayes. The neural network classifiers like GRNN and FFBPN are used for classification in MATLAB in the numeric form of data with a training and testing data ratio of 70% and 30% respectively. The accuracy of other classifiers such as SVM, DT and Naive Bayes is calculated on the nominal form of data with 80% data split. Artificial neural networks showed 85.7% of ability to discriminate and classify companies according to their competitive profile.

© 2019 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the Conference Program Chairs.

Keywords: K-Means clustering; classification models; export potential; competitiveness; data mining.

* Corresponding author. Tel.: +51920287620

E-mail address: jesussilvaUCP@gmail.com

1. Introduction

Several authors have oriented their researches to identify factors that enhance the competitive conditions for foreign trade. Thus, according to Escandón and Hurtado (2014) [1], in SMEs (Small and Medium Enterprises), the work experience of managers and owners is a fundamental variable in the results of the incursion of companies in international markets, establishing that the experience of entrepreneurs allow the growth at all levels of the organization (strategic, tactical, and operational), which affects the better conditions for success, recognize the quality of the entrepreneur (attitude to take risks) and innovation as factors of great importance for the competitiveness of companies and their positioning in international markets. Cabarcas and Paternina (2011) [2] analyze various productive factors from which they establish significant differences in the productive profiles of exporting and non-exporting companies. For its part, Smith (2005) [3] identifies environmental, organizational, managerial, strategic, and functional determinants as relevant factors in the performance of exporting companies (Correia et al., 2009) [4] and establish the effect of the quality of the relationship on performance of exports, showing that an exporter can positively influence its own performance by adopting a relationship orientation that, in turn, will affect the performance of exports.

On the other hand, the literature related to the conceptualization of the export potential is scarce. According to Obschatko and Blaio (2003) [5], the export potential of an organization is associated with the growth of exports due to its high value influenced by international, local, and technological variables among other. Paredes (2016) [6] defines it as the set of conditions that a company meets to use its strengths, weaknesses and take advantage of international market opportunities in the development of foreign trade (Lis-Gutiérrez J. et al; 2018) [7].

The use of statistical tools based on artificial neural networks for the detection of behavior patterns has thrown good results in such heterogeneous fields as the real estate market (Caridad and Ceular, 2001) [8], the biomedical area (Uberbacher and Mural, 1991) [9], or in financial markets (Olmedo and Velasco, 2007) [10]. This is the reason why this study explores the possibilities offered by the statistical tools for the recognition of business competitiveness patterns.

In this work, data mining techniques are used to classify companies according to their competitive profile and, subsequently, the Ensemble Learning to predict their export potential.

2. Materials and Proposed Method

For the development of the research, the methodology proposed by De La Hoz and López P in 2017 [11] was assumed as a reference. A group of companies from the agroindustrial sector in Colombia were taken as the object of study for a total of 272 registered in the Chamber of Commerce, to which the instrument KFM_EP (Key-Factor Measurement of the Export Potential), proposed by De la Hoz et al. (2016) [12], was applied.

The KFM_EP is a validated instrument to measure the export potential, designed from the review of the literature related to export orientation, integrating 16 key factors in export competitiveness, systematized in five dimensions: Financial (Financial Management (F1), Risk Management (F2), Financial Health (F3)), Market (Market Knowledge (M1), Foreign Trade (M2), Product Competitiveness (M3)), Learning and Growth (Information Management (AC1), Management of Knowledge (AC2), Work Environment Management (AC3), Client (Supplier Management (C1), Customer Management (C2), Requirement Management (C3)) and Internal Processes (Productivity Management (PI1), Innovation Management (PI2), Logistics Factors (PI3), and Technologies and Operations Management (PI4)).

2.1 Synthetic Minority Oversampling Technique (SMOTE) and Removal of Duplicate Rows.

SMOTE is an oversampling technique and is generally used when the data is highly imbalanced. It synthetically generates the new instances of minority class rather than traditional methods which simply replicate the minority

class or eliminate the instances from majority class (Qazi, N; 2012) [13]. New minority class instances are synthesized between the existing (real) minority samples. In Fig.1 the instances shown in red belong to minority class (4 samples) and the ones in green belong to majority class (13 samples). SMOTE is imagined to draw lines between the existing minority instances as shown below. Then, SMOTE imagines new, synthetic minority somewhere on these lines (yellow). This way the balance is maintained between the minority and majority classes.

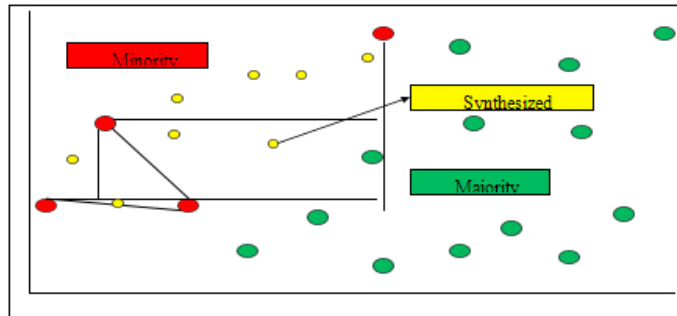


Fig.1. Working of SMOTE

On using this technique, the duplicate instances are generated, so they are eliminated at each step to get the balanced data. Table 1 demonstrates the results of applying SMOTE and remaining dataset after removing duplicate instances.

Table 1. No. of Instances after preprocessing

Total no. of instances initially = 272(R=81, NR=191)		
SMOTE(1) k=5 N=100	Total no. of instances = 353	
	R	NR
	162	191
	Total no. of samples after removing duplicate rows = 327	
	R	NR
SMOTE(2) k=5 N=100	Total no. of instances = 463	
	R	NR
	272	191
	Total no. of samples after removing duplicate rows = 356	
	R	NR
SMOTE(3) k=5 N=100	Total no. of instances = 521	
	R	NR
	330	191
	Total no. of samples after removing duplicate rows = 369	
	R	NR
Missing Data Elimination	Total no. of instances = 360	
	R	NR
	174	186

2.2 K-Means Clustering

In this, n objects are divided into k clusters based on certain attributes such that $k < n$. The main purpose is to define k centroids to each of these clusters, and to fill these clusters with the nearest items to them (Sharmila, S. and Kumar, M; 2013) [14]. K means is an iterative algorithm that tries to improve the partitioning by relocating the object from one cluster to another.

The division of instances is described in Table 2.

Table 2. Number of Instances after K-Means Clustering

Total No. of Instances	Class 1 Consolidated	Class 2 Mature	Class 3 Formation
360	71	103	186

Class 1 called "Consolidated" classifies the companies with the best competitive conditions and strengths to develop international trade. Class 2 to mature and Class 3 in formation. Next, Machine Learning techniques for prediction are defined.

2.3 Machine Learning Classifiers

2.3.1 . Generalized Regression Neural Network (GRNN);

GRNN is the variation of radial basis neural network which is used for function approximation. It is based on kernel regression networks (Kumar, G. and Malik, H; 2016) [15]. As back propagation method uses iterative training procedure, it is not required in case of GRNN. The main idea of this algorithm is to map the approximation function between the input and the target vector with the minimum error.

2.3.2 Feed Forward Back Propagation Neural Network (FFBPN)

FFBPN is one of the most commonly used algorithms in ANN and has many applications in engineering. In this work, network feed forward algorithm is used to do neural processing and pattern recall whereas the back-propagation method is used to train the neural network (Sun, G; 2008) [16], (Viloria, A and Gaitan-Angulo, M; 2016) [17]. The general architecture is the same as the ANN model.

2.3.3 Support Vector Machine (SVM)

SVM is a supervised learning technique used for non linear complex functions. It can be used for both classification and regression purposes. A model is built using the SVM training algorithm which assigns new samples to one class or the other based on the training set. There are Kernels used in SVM which check the similarity between the instances. The resultant classifier formed is now generalized enough to be used for classification of new samples (Kumar, G. and Malik, H; 2016) [15].

2.3.4 Decision Tree (DT)

The Decision Tree builds the models for classification and regression in the form of tree structure. Classes are represented through a series of yes/no questions in case of DT. Classification of the instance is the attribute specified by this node, and then going down tree branch corresponding to the value of the attribute (Qazi, N; 2012) [13]. Nodes represent the input variables and the leaves correspond to the decision outcomes (Amelec, V and Alexander, P. 2015) [18].

2.3.5 Naive Bayes

It is based on Bayes Theorem. In this, we assume that predictors are independent, i.e knowing the value of one attribute does not imply the value of other attributes. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred (Obschatko and Blaio; 2003) [5].

3. Discussion and results

Neural network classifiers such as GRNN and FFBPN are used for the classification in MATLAB in the numerical form of the data with a proportion of training and testing data of 70% and 30% respectively. The accuracy of other classifiers such as SVM, DT and Naive Bayes is calculated in the nominal data form with 80% data division. The precision results of the classifiers are presented in Table 3.

Table 3. Classification Accuracy of ML Classifiers

Graphical Representation		Classification Accuracy of ML Classifiers	
<p>Comparison of Classification Algorithms</p> <p>■ Accuracy</p>		Classifiers	Classification Accuracy
		GRNN	83.33%
		FFBPN	85.18%
		SVM	77.77%
		Decision Tree	70.83%
		Naive Bayes	72.22%

The accuracy measures such as Sensitivity, Specificity, Precision and Recall (Viloria, A and Robayo, P. 2016) [19] for all the classifiers are presented in Table 4. These measures are evaluated individually for all the classes.

Table 4. Accuracy Measures of different Classifiers

Classifier	Sensitivity	Specificity	Precision	Recall	Class
SVM	0,941	0,927	0,800	0,941	Class1
	0,421	1	1,000	0,421	Class2
	0,889	0,667	0,727	0,889	Class3
Decision Tree	1,000	0,891	0,739	1,000	Class1
	0,579	0,868	0,611	0,579	Class2
	0,639	0,778	0,742	0,639	Class3
Naive Bayes	0,941	0,909	0,762	0,941	Class1
	0,632	0,868	0,632	0,632	Class2
	0,667	0,778	0,750	0,667	Class3
GRNN	0,782	0,945	0,782	0,782	Class1
	0,782	0,934	0,750	0,782	Class2
	0,871	0,868	0,885	0,871	Class3
FFBPN	0,833	0,968	0,833	0,833	Class1
	0,821	0,909	0,741	0,821	Class2
	0,871	0,902	0,915	0,871	Class3

4. Conclusions

The developed methodology proposes data mining techniques to classify companies according to their competitive profile and subsequently the Ensemble Learning to predict their export potential and thus identify factors

on which efforts, resources, and controls should be concentrated to improve the competitive conditions and results of the company. The classifiers of neural networks as GRNN and the FFBN resulted in the most accurate techniques for the forecast made about the 272 companies in the studio.

References

- [1] Escandón, D., y Hurtado, A., Los determinantes de la orientación exportadora y los resultados en las pymes exportadoras en Colombia, *Estudios Gerenciales*, 30(133), 430–440 (2014).
- [2] Cabarcas, J., y Paternina, C., Aplicación del análisis discriminante para identificar diferencias en el perfil productivo de las empresas exportadoras y no exportadoras del Departamento del Atlántico de Colombia, *Revista Ingeniare*, 6(10), 33–48 (2011)
- [3] Smith, D., A Neural Network Classification of Export Success in Japanese Service Firms, *Services Marketing Quarterly*, 26(4), 95–108 (2005).
- [4] Correia, A., Barandas, H., y Pires, P., Applying Artificial Neural Networks to Evaluate Export Performance : A Relational Approach, *Review of International Comparative Management*, 10(4), 713–734 (2009)
- [5] Obschatko, E., y Blaio, M., El perfil exportador del sector agroalimentario argentino. Las producciones de alto valor. Estudio 1. EG.33.7. Ministerio de Economía de Argentina (2003)
- [6] Paredes, D., Elaboración del plan de negocios de exportación. Programa de Plan de Negocio, Exportador- PLANEX. Disponible en: <https://goo.gl/oTnARL> (2016)
- [7] Lis-Gutiérrez JP., Gaitán-Angulo M., Balaguera MI., Viloria A., Santander-Abril JE. (2018) Use of the Industrial Property System for New Creations in Colombia: A Departmental Analysis (2000–2016). In: Tan Y., Shi Y., Tang Q. (eds) *Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science*, vol 10943. Springer, Cham
- [8] Caridad, J. M., & Ceular, N. (2001). “Un análisis del mercado de la vivienda a través de redes neuronales artificiales”. *Estudios de economía aplicada*, (18), pp. 67-81.
- [9] Uberbacher, E. C., & Mural, R. J. (1991). “Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach”. *Proceedings of the National Academy of Sciences*, 88(24), pp.11261-11265.
- [10] Olmedo, E., Velasco, F., & Valderas, J. M. (2007). “Caracterización no lineal y predicción no paramétrica en el IBEX35”. *Estudios de Economía Aplicada*, 25(3).
- [11] De La Hoz, E., González, Á., y Santana, A., Metodología de Medición del Potencial Exportador de las Organizaciones Empresariales, *Información Tecnológica*, 27(6), 11–18 (2016)
- [12] De La Hoz, E., López P. Aplicación de Técnicas de Análisis de Conglomerados y Redes Neuronales Artificiales en la Evaluación del Potencial Exportador de una Empresa. *Información Tecnológica*. Vol. 28(4), 67-74 (2017).
- [13] Qazi, N. Effect of Feature Selection, Synthetic Minority Over-sampling (SMOTE) And Under-sampling on Class imbalance Classification. <https://doi.org/10.1109/UKSim.116> (2012)
- [14] Sharmila, S., & Kumar, M. An optimized farthest first clustering algorithm. *Nirma University International Conference on Engineering, NUiCONE 2013*, 1–5. <https://doi.org/10.1109/NUiCONE.2013.6780070> (2013)
- [15] Kumar, G., & Malik, H. Generalized Regression Neural Network Based Wind Speed Prediction Model for Western Region of India. *Procedia Computer Science*, 93(September), 26–32. <https://doi.org/10.1016/j.procs.07.177> (2016)
- [16] Sun, G., Hoff, S., Zelle, B., & Nelson, M. Development and Comparison of Backpropagation and Generalized Regression Neural Network Models to Predict Diurnal and Seasonal Gas and PM 10 Concentrations and Emissions from Swine Buildings, 0300(08) (2008).
- [17] Viloria, A., & Gaitan-Angulo, M. (2016). Statistical Adjustment Module Advanced Optimizer Planner and SAP Generated the Case of a Food Production Company. *Indian Journal Of Science And Technology*, 9(47). doi:10.17485/ijst/2016/v9i47/107371
- [18] Amelec, V., & Alexander, P. (2015). Improvements in the Automatic Distribution Process of Finished Product for Pet Food Category in Multinational Company. *Advanced Science Letters*, 21(5), 1419-1421.

- [19] Vilorio, A., & Robayo, P. V. (2016). Inventory reduction in the supply chain of finished products for multinational companies. *Indian Journal of Science and Technology*, 8(1).